# Overplacement Bias and Its Impact on Punishment and Cooperation Breakdown in Public Goods Games

Ah Chong, Leonie Ye

St. Paul's Co-educational College, Hong Kong SAR China

## Abstract

**This study investigates the relationship between biased self-assessment and punitive behaviour in social dilemmas. Using an in-person, multi-round Public Goods Game with a punishment mechanism, we examine how player overplacement, the tendency to overestimate one's contribution relative to others, predicts the use of costly punishment and its subsequent effect on group cooperation. Participants (N = 20) were assigned to five groups of four players. Before the game began, each player estimated their expected contribution level relative to their group's average. During each of the ten rounds, players decided how many chips to contribute to a public pool and later had the opportunity to punish others at a personal cost. Results indicate that players who overplaced their contributions were significantly more likely to administer punishment, particularly targeting lower contributors. Furthermore, groups with overplacers exhibited a more rapid decline in cooperation over time. These findings suggest that overplacement-driven punishment is motivated more by personal grievance than cooperative enforcement, ultimately accelerating the breakdown of collective action. This study highlights the damaging role of cognitive biases in the failure of institutions designed to sustain cooperation.**

## Keywords

**Overplacement bias, Punishment behaviour, Cooperation breakdown, Cognitive biases in economics.**

## 1. Introduction

### 1.1. Research Background and Topic

Human cooperation underpins societal progress, supporting endeavours ranging from local community initiatives to global economic systems. Despite its importance, cooperation remains vulnerable to free riding, the tendency to benefit from others' contributions without contributing oneself. The Public Goods Game (PGG), a widely used laboratory model, captures this tension: while collective welfare is maximised when all contribute to a common pool, individual incentives favour withholding contribution [1]. To address this dilemma, costly punishment, where individuals can sanction free-riders at a personal cost, has emerged as a key mechanism for maintaining cooperation [2]. Yet, punishment itself presents an anomaly: rather than always promoting cooperation, it can provoke retaliatory actions that quickly erode the very collaboration it aims to protect [3, 4].

This raises a critical question: under what conditions does punishment support cooperation, and when does it undermine it? This paper focuses on a widespread cognitive bias known as overplacement, or the "better-than-average" effect, where individuals consistently view their abilities, contributions, or moral qualities as superior to those of their peers [5, 6]. I argue that overplaced individuals enter cooperative interactions with inflated expectations of both them and others. When others fail to meet these expectations, their punitive responses are less about

enforcing norms impartially and more about expressing spite and a sense of aggrieved entitlement.

### 1.2. Research Objectives and Significance

Although theoretical connections between overconfidence and punishment have been suggested, they remain insufficiently explored in controlled experimental contexts. To address this, a straightforward, in-person PGG with punishment was conduected, measuring participants' pre-game beliefs about their relative contributions. This design enables direct testing of three hypotheses: first, that overplaced individuals deliver more punishment; second, that their punishment disproportionately targets lower contributors; and third, that the presence of overplaced players accelerates the breakdown of cooperation over repeated rounds. Ultimately, this paper contends that overplacement is a crucial predictor of costly punishment in repeated social dilemmas and that such punishment plays a central role in the rapid erosion of cooperation.

### 1.3. Research Methods and Framework

This study tests hypotheses grounded in the theoretical framework that overplacement may influence cooperation dynamics and the use of punishment.

Twenty participants were divided into five groups of four for a Public Goods Game with a punishment option. The study used a within-groups design. Overplacement was measured by comparing pre-game estimates to actual contributions. The main variables examined were punishment allocation, punishment targets, and changes in contribution levels.

## 2. Literature Review

### 2.1. Foundation and Manifestation of Overplacement

Overplacement, commonly referred to as the "better-than-average" effect, is a well-established cognitive bias where most individuals rate their own abilities, performance, or moral qualities as higher than the median of their peer group [8]. This phenomenon represents a specific type of overconfidence that is inherently social and comparative, making it particularly relevant in interactive situations, such as economic games.

The roots of overplacement lie in several psychological processes. First, the egocentric bias plays a significant role, as people focus heavily on their own efforts and skills while overlooking others' capabilities [5]. For instance, an individual is aware of their own hard work but can only speculate about a colleague's effort, leading to inflated self-assessment.

Second, overplacement serves self-enhancement needs by helping individuals maintain a positive self-image [7]. This bias often reflects motivated reasoning, where people selectively recall successes and ignore failures.

Notably, overplacement is most pronounced for easy tasks or ambiguous situations where clear feedback is lacking [6]. These conditions resemble the strategic uncertainty in repeated Public Goods Games, where contributions provide noisy information and payoffs are complex.

Consequently, individuals in cooperative social dilemmas often hold an ingrained belief that they will contribute more generously than the average participant. This biased self-view sets the stage for a pattern of behaviour that emerges when others' contributions fall short of these inflated expectations.

### 2.2. Cooperation, Punishment and Their Complications

If overplacement reveals how individuals tend to see themselves through a biased lens, then the Public Goods Game (PGG) offers a stage where the real-world consequences of this bias come into focus. The classic PGG captures a central tension between what benefits the group

and what serves each person individually: while everyone gains most when the group contributes generously, each player has a private incentive to hold back and benefit from others' efforts without pitching in [1]. This creates a familiar social dilemma, where in the absence of any reinforcing mechanisms, cooperation steadily unravels, and the group settles on a non-cooperative equilibrium.

The discovery of costly punishment, the idea that individuals might pay a price to penalise free riders more heavily, was initially celebrated as a powerful driver to sustain cooperation. Studies by Fehr and Gächter [2, 9] showed that when players could punish those who shirked, cooperation rates remained high over repeated rounds. This supported theories of strong reciprocity and altruistic punishment [10, 11], where punishment is a pro-social tool, a costly sacrifice by some players to uphold group norms that benefit everyone, regardless of future interaction.

Yet, later research painted a more complicated picture. Punishment is not always the clear-cut solution it once seemed to be. A growing body of work highlights a darker side to punishment. For one, punishment can be antisocial; players may target high contributors rather than free-riders, which is surprisingly common across many societies [3]. Those generous contributors who face punishment often cut back on their efforts, sparking a harmful cycle of decline. Furthermore, punishment can set off retaliatory spirals, where targeted players strike back, leading to destructive feuds that sap group resources and accelerate breakdowns in cooperation [13]. Finally, the impact of punishment heavily depends on the context, including the cost-benefit balance of imposing it, communication opportunities, and the rules governing its use, which can either protect or undermine collective welfare [12] [14]. This suggests that the motivation behind the punisher is the critical factor. If punishment stems from a genuine desire to enforce fair group norms, it can help sustain cooperation. But if it arises from a personal feeling of wronged entitlement, it's far more likely to trigger destructive outcomes. When punishment is driven by an inflated sense of one's own standing or grievance, it risks reflecting the very biases and social tensions that threaten collective efforts.

# 3. Methods

## 3.1. Participants and Design

Twenty participants were recruited for this study, forming five stable groups of four players each. The experiment employed a within-groups design where all participants engaged in the same experimental task. The key independent variable was the degree of overplacement, calculated from pre-game estimates and actual in-game behaviour. Dependent variables included the allocation of punishment points, punishment targets, and changes in contribution levels across rounds.

## 3.2. Experimental Procedures

Setting and Instructions: The experiment took place in a controlled environment with players physically separated to prevent communication. Participants received standardized instructions on the Public Goods Game rules, payoffs, and punishment mechanism. A practice round ensured understanding.

Pre-Game Belief Elicitation: Before the first round, participants answered: "What percentage of the average player's contribution in your group do you think your own contribution will be?" Responses were whole percentages (e.g., 150%, 80%).

Gameplay

The game comprised 5 rounds, each with the following sequence:

(1) Players received 10 chips.

(2) Players privately decided how many chips (0-10) to contribute to the communal pot.

(3) Contributions were collected, and the experimenter announced total group contribution, minimum, maximum, and average contributions.
(4) Individual preliminary earnings = chips kept + equal share of multiplied group project.
(5) Players received 5 punishment points to allocate to others; each point costs sender 1 chip and reduces recipient's earnings by 3 chips.
(6) Final earnings were calculated and recorded.
Post-Game Questionnaire: After the final round, participants completed a brief questionnaire about their reasoning behind contribution and punishment decisions.

### 3.3. Measures

Overplacement Score
Calculated as:

$$\text{Overplacement Score} = \text{Estimated Relative Contribution} - \text{Actual Relative Contribution} \quad (1)$$

$$\text{Where Actual Relative Contribution} = \left(\frac{\text{Player's Contribution}}{\text{Average Contribution of Others}}\right) \times 100\% \quad (2)$$

$$\text{Earnings were calculated as: } (10 - \text{Contribution}) + \frac{Total\ Group\ \text{Contribution} \times 1.5}{4} - (\text{Punishment Tokens Spent}) + (\text{Punishment Tokens Received}) \times 3 \quad (3)$$

Two primary measures of Punishment Behaviour:
(1) Total Punishment Sent: total punishment points allocated by a player across rounds
(2) Target-Specific Punishment: punishment points assigned based on contribution rank relative to group average
Cooperation Level: Group cooperation was measured by mean contribution per round across all groups. The rate of change in cooperation was analysed via linear regression models.

### 3.4. Hypothesis

Based on the theoretical framework of overplacement and its potential to fuel negative reciprocity, this study tests two hypotheses:
Hypothesis 1 (H1): A positive relationship exists between a player's overplacement score and their use of costly punishment. Players who overestimate their relative contribution will assign more punishment tokens throughout the game than those with accurate or underestimated contributions.
Hypothesis 2 (H2): The presence of overplaced players within a group accelerates the breakdown of cooperation. Groups with one or more highly overplaced players will show a faster decline in average contributions across 10 rounds compared to groups without highly overplaced players.

## 4. Results

### 4.1. Group Analysis

The data from the five groups revealed distinct behavioural dynamics, providing nuanced and sometimes contrasting evidence for the hypotheses. The following narratives detail the interplay between overplacement, punishment, and cooperation within each group.

### 4.1.1. Group 1: Volatility and Resilience

Group 1, comprising players AC, BC, GW, and LM, exhibited volatile contribution patterns but demonstrated notable resilience against a total collapse of cooperation. The primary overplacer, AC (score = 1.6), began with a high self-assessment yet contributed moderately (6 chips) in the initial round. Punishment was introduced in Round 2 by BC (a non-overplacer, score = 0.87), who targeted LM, the lowest contributor from Round 1. This action provoked immediate retaliation from LM directed at BC within the same round. Importantly, this punitive exchange did not escalate further. Instead, the group rebounded, achieving its highest average contribution (6.5 chips) in Round 3. In the final round, another cycle of punishment occurred between AC and BC, yet the group's mean contribution (4.5 chips) remained significantly above its lowest levels. This pattern suggests that while punishment was present, its intermittent and reciprocal nature did not precipitate a sustained decline in cooperation, indicating that limited punitive interactions may be withstood by groups without catastrophic consequences.

### 4.1.2. Group 2: Cascade of Destructive Punishment

Group 2 (DA, EB, FL, HG) provides the strongest empirical support for both hypotheses, illustrating a clear case of overplacement-driven cooperation breakdown. The group included two overplacers, most notably DA (score = 1.5), whose behaviour was instrumental. After an initially high contribution (8 chips), DA emerged as the most frequent punisher in the experiment. DA first punished HG, the lowest contributor, in Round 2, then extended punitive actions to FL in Round 3 and EB in Round 4. By Round 4, intense retaliatory punishment erupted, with EB and FL also targeting DA. This escalating conflict, overwhelmingly initiated by the highly overplaced individual, is reflected in cooperation data: mean contributions steadily declined from 6.5 chips in Round 1 to near total collapse at 0.75 chips in Round 5. This group exemplifies how an active, biased punisher can instigate a cascade of retaliation, rapidly undermining cooperative behaviour.

### 4.1.3. Group 3: Retaliation Without Breakdown

Group 3 (JT, KS, MR, PD) represents a critical exception that challenges a simplistic reading of the second hypothesis. This group had two overplacers, JT (score = 1.4) and PD (score = 1.1). JT acted on this bias in Round 2 by punishing MR, the lowest contributor from Round 1; however, this did not trigger a widespread escalation of punitive behaviour. Instead, a contained, dyadic conflict arose between JT and KS (a non-overplacer) in later rounds, likely motivated by reciprocity rather than overplacement. Notably, PD did not engage in punishment at any point. Despite these conflicts, the group maintained a relatively stable cooperative dynamic: contributions fluctuated but remained moderate to high, peaking at 7.0 chips in Round 3 and finishing at 5.5 chips. These findings suggest that limited bilateral punitive interactions can be contained without infecting the broader cooperative atmosphere, particularly when other potential punishers abstain.

### 4.1.4. Group 4: Sustained Cooperation Without Punitive Escalation

Group 4 (AN, BP, CL, DT) directly contradicts the assumption that overplacement inevitably precipitates cooperation breakdown. Despite having mild overplacers AN (score = 1.3) and DT (score = 1.05), this group was the most cooperative and stable throughout the experiment. Punishment was nearly absent, with only a single punitive token administered by DT to CL in Round 3, which did not provoke retaliation or further punitive acts. Freed from the costs and conflict associated with punishment, the group maintained high contribution levels, never dropping below a mean of 5.75 chips and concluding at 6.0 chips. This evidence indicates that overplacement bias does not necessarily translate into punitive behaviour, and in the absence of such behaviour, groups can sustain highly efficient cooperative equilibria.

### 4.1.5. Group 5: Mutual Assured Disruption

Group 5 (EM, FN, GL, HQ) illustrates a case of mutual disruption fuelled by two active overplacer-punishers. Both EM (score = 1.45) and GL (score = 1.1) engaged heavily in reciprocal punishment, frequently targeting each other, which generated a toxic interaction environment. EM punished FN in Round 3 and GL in Rounds 4 and 5, while GL punished HQ in Rounds 3 and 4 and reciprocated against EM in Round 5. This complex web of retaliation engendered persistent internal conflict, reflected in a steep decline in cooperation from an initial 6.75 chips in Round 1 to a low of 1.0 chip in Round 5. The behaviour of this group supports the second hypothesis by demonstrating that multiple overplacers acting on their bias can generate a self-reinforcing punishment cycle that effectively dismantles cooperation, even absent a single dominant instigator.

## 4.2. Overall Patterns Across Groups

An aggregate analysis across all experimental groups highlights the central role of overplacement in motivating punitive behaviour while also emphasizing that its detrimental impact on cooperation depends critically on whether this bias manifests in punitive actions.

A robust positive correlation was observed between an individual's pre-game overplacement score and the total number of punishment tokens they administered. Specifically, players with the highest overplacement scores, such as DA (1.5) and EM (1.45), were the most active punishers. Punishments were predominantly directed at members contributing below the group average, aligning with the concept of aggrieved entitlement. Nonetheless, notable exceptions were present: BC (Group 1, score = 0.87) frequently engaged in punishment despite a lower overplacement score, suggesting that factors such as reciprocal motivations may also drive punitive behaviour. Conversely, PD (Group 3, score = 1.1) refrained entirely from punishment, indicating that while overplacement predisposes individuals toward punitive actions, it does not deterministically lead to them.

Cooperation Breakdown (H2):

The presence of overplacers within a group appears necessary but not sufficient for cooperation breakdown. The critical determinant is whether overplacers convert their cognitive bias into punitive behaviour. This relationship is demonstrated by the contrasting cooperation trajectories observed between groups with active overplacer-punishers and those without. Groups featuring active overplacer-punishers (Groups 2 and 5) experienced a pronounced and rapid decline in cooperation, as confirmed by linear regression analyses revealing steep negative slopes (Group 2: $\beta = -1.56$, $R^2 = 0.87$; Group 5: $\beta = -1.48$, $R^2 = 0.76$). In contrast, groups with inactive or absent overplacer-punishers (Groups 1, 3, and 4) exhibited stable or dynamically resilient cooperation levels, characterized by significantly flatter or negligible negative slopes (Group 4: $\beta = -0.15$, $R^2 = 0.05$).

# 5. Conclusion

This study provides evidence that overplacement bias may play a meaningful role in motivating punitive behaviour within cooperative groups. However, the extent to which overplacement undermines cooperation appears to depend on whether individuals act on their biased self-assessments through punishment. Groups where overplacers actively engaged in punishment tended to experience more pronounced declines in cooperation, whereas groups with less punitive or inactive overplacers generally maintained more stable cooperative dynamics. These results suggest that the presence of overplacers alone is not sufficient to predict cooperation breakdown; rather, the behavioural follow-through in the form of punishment is a critical factor. While these findings contribute to a better understanding of cognitive biases in social dilemmas, further research is needed to explore the underlying mechanisms that influence when and why

overplacement translates into punitive actions. From an economic perspective, these insights are particularly relevant for managing cooperative behaviours in organizational and market settings, where overconfidence and punishment dynamics can impact team productivity, negotiation outcomes, and the efficiency of collective decision-making. Understanding these processes may help inform strategies to foster sustainable cooperation and mitigate costly conflicts in economic environments.

While this study provides evidence for the role of overplacement in destructive punishment, several limitations must be acknowledged. These limitations, however, serve as productive pathways for future research.

First, sample size and generalizability. The primary limitation of this study is its small sample size (N=20 participants, 5 groups). While sufficient for identifying clear behavioural patterns and effect directions, a larger and more diverse participant pool is necessary to enhance the statistical power and generalizability of the findings. Future studies should recruit a larger number of participants from varied demographics to ensure the results are not idiosyncratic to a specific population (e.g., university students or friend groups).

Second, the number of rounds and the end-game effect. The five-round design, while practical, may have intensified the "end-game effect," where players anticipate the conclusion of the game and strategically alter their behaviour (e.g., increasing defection or retaliation) in the final rounds. A longer series of rounds (e.g., 15-20) would allow researchers to distinguish more clearly between stable behavioural trends and terminal round strategies, providing a more robust test of long-term dynamics.

The findings of this study open multiple avenues for future research aimed at deepening understanding of overplacement and its influence on cooperation and punishment in social dilemmas.

First, replicating this experiment with a larger and more heterogeneous sample would enhance the robustness and generalizability of the results. Including participants from diverse cultural and socioeconomic backgrounds can uncover how contextual factors shape the prevalence and impact of overplacement biases. Cross-cultural comparisons may reveal varying propensities for overplacement and distinct norms around punishment and cooperation.

Secondly, incorporating additional psychological instruments, such as scales measuring entitlement, social dominance orientation, or aggressive tendencies, would enable a richer characterization of the motivational profiles that mediate the relation between overplacement and punishment behaviour. This could clarify whether punishment arises primarily from norm enforcement motives or from personal grievances rooted in egocentric biases.

Lastly, longitudinal designs tracking the same participants over extended periods of repeated social dilemmas would shed light on the temporal dynamics of overplacement. Such studies could assess whether repeated interaction and feedback reduce biased self-perceptions and attenuate punitive impulses, or alternatively, whether overplacement solidifies, perpetuating cycles of conflict and cooperation breakdown.

By pursuing these directions, future research can advance theoretical models of social behaviour by integrating cognitive biases with group dynamics. Ultimately, such insights could inform practical interventions and institutional designs aimed at fostering resilient cooperation and reducing conflict in collective contexts.

## References

[1] Ledyard, John O. "Public Goods: A Survey of Experimental Research." *Handbook of Experimental Economics*, edited by John H. Kagel and Alvin E. Roth, Princeton University Press, 1995, pp. 111–194.

[2] Fehr, Ernst, and Simon Gächter. "Altruistic Punishment in Humans." *Nature*, vol. 415, no. 6868, 2002, pp. 137–140.

[3] Herrmann, Benedikt, Christian Thöni, and Simon Gächter. "Antisocial Punishment Across Societies." *Science*, vol. 319, no. 5868, 2008, pp. 1362–1367.

[4] Nikiforakis, Nikos. "Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?" *Journal of Public Economics*, vol. 92, no. 1-2, 2008, pp. 91–112.

[5] Kruger, Justin. "Lake Wobegon Be Gone! The 'Below-Average Effect' and the Egocentric Nature of Comparative Ability Judgments." *Journal of Personality and Social Psychology*, vol. 77, no. 2, 1999, pp. 221–232.

[6] Moore, Don A., and Paul J. Healy. "The Trouble with Overconfidence." *Psychological Review*, vol. 115, no. 2, 2008, pp. 502–517.

[7] Alicke, Mark D., and Oksana Govorun. "The Better-Than-Average Effect." *Self and Identity*, vol. 4, no. 1, 2005, pp. 1–22.

[8] Larrick, Richard P., Kathleen A. Burson, and Jack B. Soll. "Social Comparison and Confidence: When Thinking You're Better than Average Predicts Overconfidence (and When It Does Not)." *Organizational Behavior and Human Decision Processes*, vol. 102, no. 1, 2007, pp. 76–94.

[9] Fehr, Ernst, and Simon Gächter. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review*, vol. 90, no. 4, 2000, pp. 980–994.

[10] Gintis, Herbert, et al. "Strong Reciprocity and Human Sociality." *Journal of Theoretical Biology*, vol. 206, no. 2, 2000, pp. 169–179.

[11] Boyd, Robert, et al. "The Evolution of Altruistic Punishment." *Proceedings of the National Academy of Sciences*, vol. 100, no. 6, 2003, pp. 3531–3535.

[12] Balliet, Daniel, Liesbeth Mulder, and Paul A. M. Van Lange. "Reward, Punishment, and Cooperation: A Meta-Analysis." *Psychological Bulletin*, vol. 137, no. 4, 2011, pp. 594–615.

[13] Denant-Boemont, Laurence, David Masclet, and Charles Noussair. "Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment." *Economic Theory*, vol. 33, no. 1, 2007, pp. 145–167.

[14] Egas, Martijn, and Arno Riedl. "The Economics of Altruistic Punishment and the Maintenance of Cooperation." *Proceedings of the Royal Society B: Biological Sciences*, vol. 275, no. 1637, 2008, pp. 871–878.